Morphonette: a morphological network of French

Nabil Hathout

▶ **To cite this version:**

Nabil Hathout. Morphonette: a morphological network of French. 2010. hal-00485503

HAL Id: hal-00485503
https://hal.science/hal-00485503

Preprint submitted on 20 May 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Morphonette: a morphological network of French

Nabil Hathout
Université de Toulouse
Nabil.Hathout@univ-tlse2.fr

Abstract

This paper describes in details the first version of Morphonette, a new French morphological resource and a new radically lexeme-based method of morphological analysis. This research is grounded in a paradigmatic conception of derivational morphology where the morphological structure is a structure of the entire lexicon and not one of the individual words it contains. The discovery of this structure relies on a measure of morphological similarity between words, on formal analogy and on the properties of two morphological paradigms: morphological derivational families and morphological derivational series.

1 Paradigmatic derivational morphology

The starting points of this research are the fundamental ideas of lexeme-based morphology (Aronoff, 1994): only lexemes are signs (i.e. atomic units); affixes are merely phonologial marks; the construction of the meaning and of the form of a derived word are distinct processes. It is grounded in a conception of derivational morphology where words do not have a morphological structure and where this structure is a level of organization of the lexicon. This organization is based on the semantic, formal and categorical relations that hold between the words memorized in the lexicon (Bybee, 1995). Among these relations, analogies play a prominent role because they allow the emergence of the morphological paradigms. An analogy is a quaternary relations $a : b :: c : d$ that holds between the members of a quadruplet (a, b, c, d) such that a is to b as c is to d. Morphological derivational analogies holds between the members of two types of paradigms : morphological derivational families and morphological derivational series. This can be illustrated with an analogy such as *duplication* : *duplicateur* :: *unification* : *unificateur*[1] where we can see that *duplication* and *duplicateur* belong to the same derivational family and that it goes the same for *unification* and *unificateur*. This conception enables us to redefine the morphological analysis task, which aims to make explicit the morphological paradigms of the lexicon instead of decompose the individual words into morphemes. This organization is illustrated in figure 1. The analysis of a given word then consists in identifying its position in the morphological structure of the lexicon. For instance, the word *rectificateur* 'recitifier' is not analyzed as in (1) but as a member of the derivational family which contains *rectifiable*, *rectifier* 'rectify', *rectifieur* 'recitifier', *rectification*, *rectificatif* 'corrective', etc. and of the derivational series which contains *certificateur* 'certifier', *fructificateur* 'which bears fruits', *modificateur* 'modifier', *sanctificateur* 'sanctifier', etc. These two sets can be seen as the morphological coordinates of *rectificateur*.

[1] 'duplication', 'duplicator', 'unification', 'unifier'

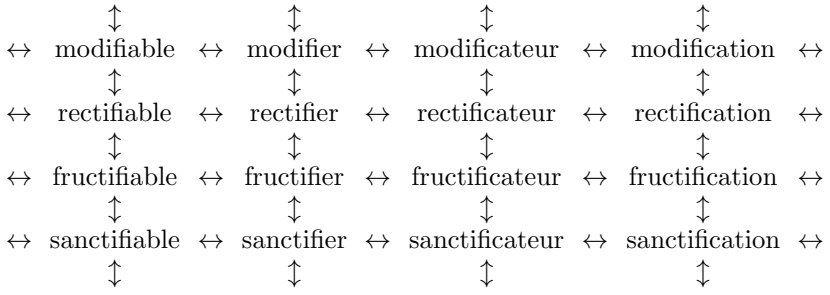
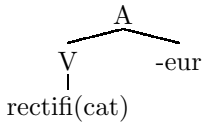


Figure 1: The morphological network of the French lexicon is made up of derivational families and derivational series. Families and series are connected by morphological analogies.

(1)


The objective of the present research is twofold: first, we propose a radically lexeme-based method of morphological analysis capable of providing the morphological derivational structure of the lexicon; second, we have computed this structure for a significant fragment of a large-coverage lexicon of French. This resource, Morphonette, will soon be made available to the public.

A morphological network solves several problems posed by the morphematic approach such as the treatment of words such as *concevoir* 'conceive', *décevoir* 'deceive', *percevoir* 'perceive', *recevoir* 'receive' or *consister* 'consist', *désister* 'desist', *persister* 'persist', *résister* 'resist' where it is difficult to determine the status to the *con-*, *dé-*, *per-*, *re-*, *-cevoir* or *-sister* sequences. The paradigmatic approach is also capable of bringing words such as *furieux* 'furious' and *curieux* 'curious' into the same lexical derivational series despite the fact that *furieux* has a derivational base, *furie* 'fury', while the current lexicon of French contains no word that could serve as a base to *curieux*. The dissociation of the construction of meaning and form allows us to easily treat allomorphy, suppletion and phenomena such as interfixation that one observes in *goutte* 'drop' → *gouttelette* 'droplet' or *triste* 'sad' → *tristounet* 'gloomy' described by Plénat (2005).

The network illustrated in figure 1 is actually made up of analogies. For instance, *fructificateur*:*fructification* participates in analogies with *modificateur*:*modification*, *rectificateur*:*rectification*, *sanctificateur*:*sanctification*. Similarly, *fructificateur*:*rectificateur* forms analogies with *fructifier*:*rectifier*, *fructification*:*rectification*, *fructifiable*:*rectifiable*. Gathering all theses analogies poses a serious problem of complexity. For instance, for a lexicon of 97 010 entries such as the *Trésor de la Langue Française* (TLF) word list, the number of quadruplets to be tested is on the magnitude of 10^{19}. This number is theoretically 10^{20} but it can be divided by 8 by taking advantage of the permutations described in (2) where $\mathcal{L}$ is a set of representations of the lexical units.

$$\forall (a,b,c,d) \in \mathcal{L}^4, a:b::c:d \Rightarrow a:c::b:d \wedge b:a::d:c \wedge \\ b:d::a:c \wedge c:a::d:b \wedge c:d::a:b \wedge d:a::c:b \wedge d:c::a:b \quad (2)$$

For the construction of the Morphonette network, we have used the phonological representations of the TLF headwords instead of their written forms, so reducing the size of the lexicon to 83 082 entries and the number of quadruplets to be checked to $6 \cdot 10^{18}$.

The solution we adopted for the complexity problem consists in using the measure of morphological similarity proposed by Hathout (2008). This measure enables us to select for a given

entry w the words that are most likely to form analogies with w, namely the members of the derivational family and series of w (see section 2). The second problem we have had to solve is the actual verification of the analogies. We have used the same algorithm as Hathout (2008). Inspired by the one of Lepage (1998), this algorithm allows us to check whether a formal analogy holds between four words without having to cut them into morphemes. Notice that this algorithm may exceptionally fail to find some analogies. Another algorithm, proposed by Stroppa (2005), does not suffer from this drawback. However, we did not use it because its complexity is in $o(n^4)$ while the former has a complexity in $o(n^2)$ and because these exceptional failures are largely compensated by the number and the redundancy of the collected analogies. The construction of Morphonette poses a third problem, namely the exclusion of the formal analogies that are not morphologically valid such as *constituable* : *constant* :: *restituable* : *restant*[2]. We relied on the structure of the morphological graph to eliminate them, namely on the fact that series contain large numbers of words, that they are clusters with highly connected members and that series are connected to each others by large numbers of edges which form analogies. Notice that the series of the lexicon too form a cluster.

The remainder of the paper is organized as follows. In Section 2, we present the measure of formal similarity and the morphological neighborhoods where the analogies are looked for. Section 3 outlines the verification of the formal analogies. In Section 4, we describe in detail the bootstrapping algorithm we have used for the construction of this first version of Morphonette. The resource is presented in Section 5. Section 6 discusses some related works and finally, Section 7 offers a short conclusion.

2 Morphological similarity

We have used the measure of morphological similarity proposed by Hathout (2008) for the construction of Morphonette. This measure brings closer the words that share large numbers of very specific formal and semantic features: the more features the words share and the more specific these features are, the closer they are. The measure is calculated by means of a bipartite graph where the words are connected to their features. The neighbors of a word w are identified by spreading an activation initiated at the vertex that represents w. First, the activation is uniformly spread toward the features of w. Then, in the second step, the activation located on the features is uniformly spread toward the words that possess these properties. The level of activation obtained by a word x after the propagation is an estimation of the morphological relatedness between w and x. The spreading is simulated by means of a classical random walk algorithm, that is by multiplying the stochastic adjacency matrix of the bipartite graph.

The measure originally proposed by Hathout (2008) uses both formal and semantic properties, the latter being n-grams of words extracted from the TLF definitions. We did not retain them here because they are not informative enough. Another difference with Hathout (2008) is the use of phonetic transcriptions instead of word forms. We have used the LIA_PHON phonetizer of Béchet (2001) in order to transcribe the word forms into sequences of phonemes in Mbrola format. Each phoneme is encoded as two characters as shown in the examples in (3).

(3)

constant	kkonssttan
constituable	kkonssttiittuuaabbllee
restant	rraissttan
restituable	rraissttiittuuaabbllee

The beginning and the end of the words are marked by `##`. The morphological similarity is then

[2] 'constitutable', 'constant', 'restitutable', 'remaining'

fructifier fructifiant fructificateur fructification fructifiant fructifère *sanctifier rectifier présanctifier* **fructivore** fructidorien fructidorienne fructidoriser fructidor **fructueusement fructueux fructuosité fructose** obstructif constructif instructif désobstructif destructif instructif autodestructif **usufructuaire infructueusement** sanctifiant sanctifiable rectifieuse rectifieur rectifiant rectifiable *transsubstantifier substantifier stratifier cimentifier certifier savantifier refortifier ratifier présentifier pontifier plastifier notifier nettifier mortifier mythifier mystifier quantifier*

Figure 2: The 50 nearest neighbors of *fructifier* 'bear fruit'. The members of the derivational family are in bold face and the ones of the derivational series are in italic.

estimated by associating with each word the set of all the sequences of 3 phonemes or more. For instance, the sequences which describe the word *constant* are presented in (4).

(4) `##kkon kkonss onsstt sstan ttan##`
`##kkonss kkonsstt onsstan sstan##`
`##kkonsstt kkonsstan onsstan##`
`##kkonsstan##`

Figure 2 presents the nearest neighbors of *fructifier* 'bear fruit'. If we omit *sanctifier* 'sanctify', *rectifier* and *présanctifier* 'presanctify', we see that the members of the derivational family of *fructifier* all appear at the beginning of the list and that the end gathers the members of its derivational series.

3 Formal analogy

The measure of morphological similarity enables us to determine a morphological neighborhood for each word w. This neighborhood gathers a large part of the members of the derivational family and series of w. These members are precisely the ones with which w can form morphological analogies. In this way, we can reduce drastically the search space for analogies, as proposed in Hathout (2008). For instance, if we limit the search to the 100 first neighbors of each word, the number of quadruplets to be checked for a lexicon of 83 082 entries drops to 10^{10}. This number can be further reduced by using two heuristics based on the properties (5) and (6).

$$\forall(a,b,c,d) \in \mathcal{L}^4, a : b :: c : d \Rightarrow l(a) - l(b) = l(c) - l(d) \tag{5}$$

where $l(x)$ is the number of phonemes in x.

$$\begin{aligned} &\forall(a,b,c,d) \in \mathcal{L}^4, a : b :: c : d \Rightarrow \\ &\quad (c(a) = c(b) \wedge c(c) = c(d)) \vee (c(a) = c(c) \wedge c(b) = c(d)) \end{aligned} \tag{6}$$

where $c(x)$ is the morphosyntactic tag of x. Morphonette uses the Grace tag set (Rajman et al., 1997). These heuristics divide the total number of quadruplets to be checked by 50. $2 \cdot 10^8$ quadruplets have therefore been checked and $4.2 \cdot 10^6$ formal analogies have been collected. In order to further improve the quality of these analogies, we have only kept the ones where a formal analogy also holds for the written forms. This additional condition eliminates phonetic analogies such as *paissant* : *abaissant* :: *paye* : *abeille*[3]. The number of analogies actually used for the construction of the first version of Morphonette is $3.9 \cdot 10^6$. The set of these analogies is closed under the permutations described in (2). Let $\mathcal{A}$ be this set.

[3] 'grazing', 'lowering', 'pay, 'bee'

The analogies in $\mathcal{A}$ have been found by using the same technique as the one of Hathout (2008) which consists in computing an analogical signature for each of the pairs of words (a, b) and (c, d) of a quadruplet (a, b, c, d). The analogical signature of a pair of words (a, b) describes a path in their edit lattice, that is a sequence of string edit operations. (a, b, c, d) is an analogy if the two signatures are identical. This method fails to detect some analogies such as (7).[4]

(7) *do* : *doable* :: *read* : *readable*

These failures being exceptional and the analogies highly redundant, it is always possible to recover the relations $a : b$ and $c : d$ and then the entire analogy $a : b :: c : d$. Notice that the algorithm of Stroppa (2005) is able to identify (7), but it has a complexity in $o(n^4)$. It is obviously not adapted to our needs given the number of quadruplets we have to check.

4 Morphological network

Morphonette has been constructed by using a bootstrapping algorithm. We first selected an initial seed, $\mathcal{M}_0$, composed of the most reliable morphological relations and then complemented it iteratively with relations induced by $\mathcal{M}_0$. More specifically, the $3.9 \cdot 10^6$ collected analogies were used to define a weighted graph $\mathcal{G} = (V, E, w)$ where V is a set of vertices, namely the set of the headwords of the TLF, $E = \{(a, b) \in V \times V / \exists a : b :: c : d \in \mathcal{A}\}$ a set of edges and $w : E \to \mathbb{N}$ a weight function such that $\forall e \in E, w(e) = |\{a : b :: c : d \in \mathcal{A} / (a, b) = e\}|$. $\mathcal{G}$ being build from formal analogies, the words represented by the vertices are mainly connected to members of their derivational families on one hand and to members of their derivational series on the other. The main objective of the construction of Morphonette is to set apart these two types of relations and to select a set of relations with almost no error. This is because $\mathcal{A}$ contains formal analogies such as *destructeur* : *structural* :: *descripteur* : *scriptural*[5] which induce morphologically invalid edges, namely *destructeur*:*structural* and *descripteur*:*scriptural*.

The relations between members of the same family and members of the same series can be partially set apart on the basis of the categorical features of the words: two words that belong to the same series have identical morphosyntactic tags. As a result:

$$\forall a : b :: c : d \in \mathcal{A}, c(a) \neq c(b) \Rightarrow \phi(a, b) \wedge \phi(c, d) \wedge \sigma(a, c) \wedge \sigma(b, d) \tag{8}$$

$$\forall a : b :: c : d \in \mathcal{A}, c(a) \neq c(c) \Rightarrow \phi(a, c) \wedge \phi(b, d) \wedge \sigma(a, b) \wedge \sigma(c, d) \tag{9}$$

where $\phi(x, y)$ is true iff x and y belong to the same derivational family and $\sigma(x, y)$ is true iff x and y belong to the same derivational series. However, this criterion does not allow us to type the edges of analogies where $c(a) = c(b) = c(c) = c(d)$ such as *développeur* : *développement* :: *enveloppeur* : *enveloppement*[6] which holds between four masculine singular nouns. The statements (8) and (9) can be used to define a type function τ of the analogies in $\mathcal{A}$:

$$\tau(a : b :: c : d) = \begin{cases} \text{f} & \text{if } c(a) \neq c(b) \\ \text{s} & \text{if } c(a) \neq c(c) \\ \text{u} & \text{othewise} \end{cases} \tag{10}$$

We can then define the subset of E made up of the edges which connect words which may be in the same family:

$$\mathcal{F} = \{(a, b) \in E / \exists a : b :: c : d \in \mathcal{A}, \tau(a : b :: c : d) \in \{\text{f}, \text{u}\}\} \tag{11}$$

[4] We thank Philippe Langlais who pointed out this problem to us.

[5] 'destructor', 'structural', 'descriptor', 'scriptural'

[6] 'developer', 'development', 'enveloper', 'envelopment'

The partial typing of the edges in $\mathcal{G}$ can be refined on the basis of two structural characteristics of the morphological network. These characteristics allows us to select a subgraph of $\mathcal{G}$ with the most reliable morphological relations only:

(12) Derivational series are large sets.

(13) Derivational series are clusters.

The characteristic (12) allows us to identify reliable family relations. This is because two words a and b which belong to the same family normally participate to one analogy with each of the members of the series of a and of b. Series being large sets, the weight $w(e)$ of an edge (a, b) connecting members of the same family is normally high. In other words, the number of analogies which contain a given edge can be used identify the ones which reliably connect members of the same family. For instance, a threshold of 10 can be used to select a set which only contains family edges. Let $\mathcal{F}_0 = \{e \in \mathcal{F}/w(e) \geq 10\}$ be this set. We can then rely on $\mathcal{F}_0$ to identify relations between words which belong to the same series:

$$\forall a : b :: c : d \in \mathcal{A}, (a, b) \in \mathcal{F}_0 \Rightarrow \sigma(a, c) \wedge \sigma(b, d) \tag{14}$$

$\mathcal{F}_0$ can therefore be used to extract a subgraph $\mathcal{G}_0$ from $\mathcal{G}$ composed with serial relations induced by the reliable familial relations in $\mathcal{F}_0$:

$$\mathcal{S}_0 = \{(a, c) \in E/\exists (a, b) \in \mathcal{F}_0 \text{ and } \exists a : b :: c : d \in \mathcal{A}\} \tag{15}$$

$$\mathcal{G}_0 = \mathcal{F}_0 \cup \mathcal{S}_0 \tag{16}$$

The structure we get is actually more complex. This is because one word c can belong to the series of a word a when a is in a relation with a member b of its family but not belong to the series of a when a is in a relation with another member b'. For instance, *artificiel* 'artificial' belongs to the same series as *officiel* 'official' and *troisième* 'third' when it is in a relation with *artificiellement* 'artificially' but it is only in the same series as *officiel* when in a relation with *artificialiser* 'artificialize'. In the first case, *artificiel*:*artificiellement* forms analogies with *officiel*:*officiellement* 'officially' and *troisième*:*troisièmement* 'thirdly'; in the second, *artificiel*:*artificialiser* only forms an analogy with *officiel*:*officialiser* 'officialize' but none with a pair having *troisième* as its first member. In other words, each entry belong to as many distinct sub-series as there are members in its family. Thus, the morphological structure of the lexicon consists in a set of **filaments** of the form $(a, b, series(a, b))$ where a is an entry, b a member of its family and $series(a, b) = \{c \in V/\exists a : b :: c : d \in \mathcal{A}\}$ the sub-series of a when we consider its relation with b. Actually, the filaments of an entry a are just a representation of the set of the analogies which contain a.[7] Filaments are illustrated in figure 3.

The characteristic (13) is then used to enhance the selection of the most reliable edges in $\mathcal{G}$ starting from the most central serial relations. Even if almost all the familial relations in $\mathcal{F}_0$ are correct, we need to eliminate the ones that may yield errors when the initial seed is extended, and especially the ones that connect distinct families. These connections primarily concern compounds such as *zoophilie* 'zoophilia' which belong to the family of *zoologie* 'zoology' *zoophobie* 'zoophobia', etc. and to the one of *anthropophilie* 'anthropophilia', *bibliophilie* 'bibliophilia', etc. depending on whether we consider its radical is *zoo* or *philie*. In this case, we eliminate the relation between *zoophilie* and *anthropophilie* by relying on the fact that *zoophilie* has predominantly

[7] Let us notice that filaments could be defined in a dual manner from the derivational series. In this case, a filament of an entry a is a triplet $(a, b, family(a, b))$ where b is a member of the series of a and $family(a, b)$ is the sub-family of a when we consider its relation with b. Both types of filaments being equivalents, we have used the first one because it yields a more compact description of the graph.

gazouillarde *gazouillage*
cafouillarde grenouillarde vasouillarde

gazouillarde *gazouillement*
braillarde geignarde grognarde

gazouillarde *gazouiller*
citrouillarde douillarde grenouillarde rouillarde souillarde vadrouillarde vasouillarde

Figure 3: Three filaments of the entry *gazouillarde* 'twittering female'

words ending in *-philie* in its series and that these words do not have words starting with *zoo-* in their series. Put differently, the words starting with *zoo-* are not well connected within the central cluster of the series of *zoophilie*. We classically measure the clustering coefficient of a word c within the series of a word a by the ratio of the number of triangles to the number of triples which contain the edge (a, c) (Watts & Strogatz, 1998). Let $s_0(a) = \{c \in V / (a, c) \in S_0\}$ be the series of a. Then the number of triples formed by a and one word $c \in s_0(a)$ is $|s_0(a)| - 1$. The number of triangles that a word $c \in s_0(a)$ form with other members of $s_0(a)$ is $|(s_0(a) \setminus \{c\}) \cap (s_0(c) \setminus \{a\})|$. A threshold of 0.66 has been used for the construction of Morphonette. It allows us to reduce the series to their most central clusters. For series $s_0(a)$, this cluster can be defined as in (17).

$$s'_0(a) = \{c \in s_0(a) / \frac{|(s_0(a) \setminus \{c\}) \cap (s_0(c) \setminus \{a\})|}{|s_0(a)| - 1} \geq 0.66\} \tag{17}$$

This reduction is then used to remove from $\mathcal{F}_0$ the edges (a, b) such that $series(a, b) \cap s'_0(a) = \emptyset$. The resulting graph is the initial seed $\mathcal{M}_0$.

$\mathcal{M}_0$ is then iteratively extended until a fixed-point is reached. At step i, we generate all the formal analogies induced by the transitive closures of the families of $\mathcal{M}_i$. These analogies $a : b :: c : d$ consists of to pairs (a, b) and (c, d) such that $\exists (t_1, t_2) \in \mathcal{T}_i \times \mathcal{T}_i, (a, b) \in t_1 \times t_1$ and $(c, d) \in t_2 \times t_2$ where $\mathcal{T}_i$ is the transitive closure of the families of $\mathcal{M}_i$. We then reduced the graph induced by these analogies to its intersection with $\mathcal{G}$ and added this extension to $\mathcal{M}_i$ in order to yield $\mathcal{M}_{i+1}$. We actually impose to the extension an additional condition: for $i \geq 2$, only the filaments with a sub-series of 5 words or more are kept. The fixed-point is reached in 8 iterations. The Morphonette network is the constructed by merging $\mathcal{M}_8$ with $\mathcal{G}_0$.

5 Morphonette 0.1

This first version of Morphonette comprises 29 310 entries and 96 107 filaments, and therefore the same number of familial relations. The number of distinct families has not been computed. The network contains 1 160 098 serial relations, that is 12 per filament in average. These numbers can be compared with the ones of $\mathcal{G}$, the graph from which this network has been extracted. $\mathcal{G}$ comprises 75 832 entries, 816 922 filaments (that is 10 per entry in average, against only 3 in Morphonette) et 2 343 059 serial relations (that is less than 3 per filament). Morphonette therefore already covers about 40% of the entries of the lexicon. Figure 3 presents an excerpt of this resource consisting of three filaments of the noun *gazouillarde* 'twittering female'.

A first estimation of the quality of Morphonette has been performed by manually checking 200 filaments randomly extracted from the network. Only one erroneous relation has been found between *pension* and *pensif* 'pensive' which puts the precision above 99%, if confirmed by a more thorough evaluation. Even if *pension* and *pensif* are etymologically related, there is nowadays

no semantic relation between them. However, *pension* and *pensif* participate to a large number of formal analogies which wrongly put *pension* in the extended series of deverbal nouns ending in *-ion*. The loss of the semantic relation between *pension* and *pensif* can only be detected on the basis of semantic information. But Morphonette 0.1 has been constructed only from the formal properties of the TLF headwords.

Morphonette 0.1 also contains some errors due to formal accidents such as the relation between *dégrimer* 'remove the make-up' and *dégression* 'degression' which belongs, from a formal point of view, to the series of *déprimer*:*dépression*[8], *comprimer*:*compression*[9], etc. Once again, the use of semantic knowledge should be the best way to find out and eliminate this type of errors. Another line of investigation would be to generalize the notion of analogy to sets of three pairs or more in order to determine the invariants of the sub-series.

Another difficulty we will have to address is the treatment of homonyms and homographs. For instance, the four meanings of *fraise* ('strawberry', 'mesentery', 'ruff', 'drill') induce four distinct derivational families even if the three latter meanings are etymologically related. In Morphonette 0.1 these families are confused. We will use the homonyms numbers in the TLF entries and the semantic information present in the definitions to separate them in future versions of Morphonette.

6 Related Works

From a theoretical point of view, this work belongs to a framework related to the Network Morphology of Bybee (1995), to the Surface-to-Surface Morphology of Burzio (2002), and to emergentist approaches of Aronoff (1994), Albright (2002) or Goldsmith (2006).

The construction of Morphonette uses a bootstrapping algorithm in order to extend an initial reliable seed. This technique has often also been used in computational morphology, for instance by Goldsmith (2006) or by Bernhard (2006). However, our method differs from these ones because it is fully lexeme-based and does not make use of morpheme nor contain any representation of them. Morphological regularities emerge from a very large set of analogies. Gathering of this set is one of contributions of the work presented in this paper. It was made possible through the use of the measure of morphological similarity of Hathout (2008). This measure was inspired by work on small words by Gaume et al. (2002). Our method is also close to the ones of Yarowsky & Wicentowski (2000) and Baroni et al. (2002) where the words are not decomposed into morphemes. Both make use of string edit distance to identify formal similarity between words. Our work is also close to the one by Stroppa & Yvon (2005), Langlais et al. (2009) and Lavallée & Langlais (2009) who use formal analogies to analyze words morphologically and to translate them.

The Morphonette network could also be compared to the morphological families constructed by Xu & Croft (1998), Gaussier (1999) or Bernhard (2009) among others. It is also very close to Polymots, a manually-constructed morphological lexicon (Gala et al., 2010). Polymots and Morphonette are complementary since the former primarily contains short words while the latter mainly contains long words because of the criteria we have used to select the morphological relations.

With respect to these related works, the main contribution of Morphonette is first the generation of a collection of more than 4 millions formal analogies and the exploitation of the structural properties of the morphological graph in order to set apart the familial and the serial relations.

[8] 'depress', 'depression'
[9] 'compress', 'compression'

7 Conclusion and directions for further research

We have presented in this paper Morphonette, the first morphological network of French. This network is constructed without decomposition of the words into morphemes. The method we have used rely on the structural properties of a graph of morphological relations build from a collection of almost 4 millions formal analogies. Morphonette is made up of filaments which are composed of an entry, a member of its derivational family and derivational sub-series of similar words. It allows us to redefine the morphological analysis task which does not aim to decompose words into morphemes but aims to identify their derivational families and series by means of a set of filaments.

Morphonette will soon be distributed under Creative Commons licence. A thorough evaluation of its relations will also be carried out shortly. A second version of this resource will be developed by designing a measure of semantic relatedness able to differentiate between homonyms, to spot out the formal accidents and to identify allomorphy and suppletion. This measure will be based on the relations in Morphonette 0.1 which will be used to select the semantic properties and relations which are the most informative from a morphological point of view.

References

Albright, A. (2002). *The identification of bases in morphological paradigms.* PhD thesis, University of California, Los Angeles.

Aronoff, M. (1994). *Morphology by Itself. Stem and Inflectional Classes.* Linguistic Inquiry Monographs. Cambridge, Mass.: MIT Press.

Baroni, M., Matiasek, J., & Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-2002* (pp. 48–57). Philadelphia, Penn.: ACL.

Bernhard, D. (2006). Automatic acquisition of semantic relationships from morphological relatedness. In *Advances in Natural Language Processing, Proceedings of the 5th International Conference on NLP, FinTAL 2006*, volume 4139 of *Lecture Notes in Computer Science* (pp. 121–132). Berlin / Heidelberg: Springer Verlag.

Bernhard, D. (2009). Morphonet: Exploring the use of community structure for unsupervised morpheme analysis. In *Working Notes for the MorphoChallenge at CLEF 2009* Corfu, Greece.

Burzio, L. (2002). Surface-to-surface morphology: when your representations turn into constraints. In P. Boucher (Ed.), *Many Morphologies* (pp. 142–177). Somerville, Mass.: Cascadilla Press.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, 10(5), 425–455.

Béchet, F. (2001). LIA_PHON : un système complet de phonétisation de textes. *Traitement automatique des langues*, 42(1), 47–67.

Gala, N., Rey, V., & Zock, M. (2010). A tool for linking stems and conceptual fragments to enhance word access. In *Proceedings of the seventh international conference on Language Resources and Evaluation (LREC 2010)* La Valette, Malta.

Gaume, B., Duvigneau, K., Gasquet, O., & Gineste, M.-D. (2002). Forms of meaning, meaning of forms. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(1), 61–74.

Gaussier, E. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing* College Park, MD.

Goldsmith, J. (2006). An algorithm for the unsupervised learning of morphology. *Natural Language Engineering*, 12(4), 353–371.

Hathout, N. (2008). Acquisition of the morphological structure of the lexicon based on lexical similarity and formal analogy. In *Proceedings of the Coling workshop Textgraphs-3* (pp. 1–8). Manchester: ACL.

Langlais, P., Yvon, F., & Zweigenbaum, P. (2009). Improvements in analogical learning: Application to translating multi-terms of the medical domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)* (pp. 487–495). Athène.

Lavallée, J.-F. & Langlais, P. (2009). Morphological acquisition by formal analogy. In *Working Notes for the MorphoChallenge at CLEF 2009* Corfu, Greece.

Lepage, Y. (1998). Solving analogies on words: An algorithm. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and of the 17th International Conference on Computational Linguistics*, volume 2 (pp. 728–735). Montréal.

Plénat, M. (2005). *Rosinette*, *cousinette*, *putinette*, *starlinette*. Décalage, infixation et épenthèse devant *-ette*. In I. Choï-Jonin, M. Bras, A. Dagnac, & M. Rouquier (Eds.), *Questions de classification en linguistique : méthodes et descriptions. Mélanges offerts au Professeur Christian Molinier* (pp. 275–298). Berne: Peter Lang.

Rajman, M., Lecomte, J., & Paroubek, P. (1997). *Format de description lexicale pour le français. Partie 2 : Description morpho-syntaxique*. Technical report, EPFL & INaLF. GRACE GTR-3-2.1.

Stroppa, N. (2005). *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. Thèse de doctorat, École nationale supérieure des télécommunications, Paris.

Stroppa, N. & Yvon, F. (2005). An analogical learner for morphological analysis. In *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)* (pp. 120–127). Ann Arbor, MI: ACL.

Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393, 440–442.

Xu, J. & Croft, W. B. (1998). Corpus-based stemming using co-occurrence of word variants. *ACM Transaction on Information Systems*, 16(1), 61–81.

Yarowsky, D. & Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the Association of Computational Linguistics (ACL-2000)* (pp. 207–216). Hong Kong.